计算机视觉经典综述 / 02

深度学习

Deep learning

Yann LeCun • Yoshua Bengio • Geoffrey Hinton

Nature 521, 436–444 (2015) DOI: 10.1038/nature14539

全文中文译本。正文、图注和作者评述均已翻译，原图、公式及参考文献编号保留。文中“目前”“近期”“未来”沿用 2015 年的发表语境。

作者单位 2

摘要 2

引言 2

监督学习 3

用反向传播训练多层架构 4

卷积神经网络 6

利用深度卷积网络理解图像 7

分布式表示与语言处理 9

循环神经网络 10

深度学习的未来 11

致谢 11

作者信息 12

参考文献 12

术语表 17

作者单位

[1] Facebook 人工智能研究院，美国纽约州纽约市 Broadway 770 号，邮编 10003。[2] 纽约大学，美国纽约州纽约市 Broadway 715 号，邮编 10003。 [3] 蒙特利尔大学计算机科学与运筹学系，André-Aisenstadt 楼，Centre-Ville STN 邮政信箱 6128，加拿大魁北克省蒙特利尔，邮编 H3C 3J7。[4] Google，美国加利福尼亚州山景城 Amphitheatre Parkway 1600 号，邮编 94043。[5] 多伦多大学计算机科学系，加拿大安大略省多伦多市 King's College Road 6 号，邮编 M5S 3G4。

摘要

深度学习使由多个处理层组成的计算模型能够学习具有多个抽象层次的数据表示。这些方法显著提升了语音识别、视觉对象识别、目标检测，以及药物发现和基因组学等许多领域的最佳性能。深度学习利用反向传播算法，指出机器应如何改变内部参数，使各层能够根据前一层的表示计算本层表示，从而发现大型数据集中的复杂结构。深度卷积网络使图像、视频、语音和音频处理取得了突破，而循环网络则为文本、语音等序列数据的处理开辟了新途径。

引言

机器学习技术支撑着现代社会的许多方面：从网络搜索、社交网络内容过滤，到电子商务网站的推荐；它也越来越多地出现在相机、智能手机等消费产品中。机器学习系统用于识别图像中的对象、将语音转写为文字、把新闻、帖子或商品与用户兴趣相匹配，以及选择相关的搜索结果。这些应用正越来越多地采用一类称为深度学习的技术。

传统机器学习技术直接处理自然数据原始形式的能力有限。几十年来，构建模式识别或机器学习系统，需要精心的工程设计和丰富的领域知识，以设计一个特征提取器：它把原始数据（例如图像的像素值）变换为合适的内部表示或特征向量，再由学习子系统——通常是分类器——检测或分类输入中的模式。

表示学习是一类允许机器直接接收原始数据，并自动发现检测或分类所需表示的方法。深度学习是具有多个表示层次的表示学习方法：它通过组合简单但非线性的模块来获得这些层次，每个模块把一个层次的表示（从原始输入开始）变换为更高、稍微更抽象的表示。组合足够多的这类变换，便能学习非常复杂的函数。对于分类任务，高层表示会增强输入中对判别重要的方面，并抑制无关变化。例如，图像以像素值数组的形式输入；第一层表示中学到的特征通常表示图像中特定方向、特定位置的边缘是否存在。第二层通常通过发现边缘的特定排列来检测局部模式，而不受边缘位置微小变化的影响。第三层可以把局部模式组合成更大的结构，对应于熟悉对象的部件；随后各层则把这些部件的组合作为对象进行检测。深度学习的关键在于，这些层次的特征不是人工工程师设计的，而是使用通用学习过程从数据中学到的。

深度学习正在解决一些人工智能研究界多年来尽最大努力仍难以解决的问题，并取得重大进展。事实证明，它非常擅长发现高维数据中的复杂结构，因此适用于科学、商业和政府事务的许多领域。除了刷新图像识别[1–4]和语音识别[5–7]的纪录，它在预测潜在药物分子的活性[8]、分析粒子加速器数据[9,10]、重建脑神经回路[11]，以及预测非编码 DNA 中的突变对基因表达和疾病的影响[12,13]等任务上，也超过了其他机器学习技术。或许更令人惊讶的是，深度学习在自然语言理解[14]的多种任务上取得了极有前景的结果，尤其包括主题分类、情感分析、问答[15]和语言翻译[16,17]。

我们认为，深度学习在不久的将来还会取得更多成功，因为它只需要很少的人工工程设计，因此能够很容易地利用可用计算量和数据量的增长。目前正在为深度神经网络开发的新学习算法和新架构，只会进一步加快这一进程。

监督学习

无论是否采用深度模型，机器学习最常见的形式都是监督学习。设想我们要构建一个系统，将图像分类为包含房屋、汽车、人或宠物。首先，我们收集一个由房屋、汽车、人和宠物图像组成的大型数据集，每张图像都标注所属类别。在训练过程中，机器看到一张图像，并输出一个分数向量，每个类别对应一个分数。我们希望正确类别的分数在所有类别中最高，但在训练之前，这通常不会发生。我们计算一个目标函数，衡量输出分数与期望分数模式之间的误差（或距离）。随后，机器修改内部可调参数，以减小这一误差。这些可调参数通常称为权重，是一些实数，可以视为定义机器输入—输出函数的“旋钮”。典型的深度学习系统可能拥有数亿个这样的可调权重，并利用数亿个带标签的样本进行训练。

为了恰当地调整权重向量，学习算法计算一个梯度向量。对于每个权重，该向量都指出：如果权重增加一个微小量，误差会增加或减小多少。随后，沿梯度向量的相反方向调整权重向量。

对全部训练样本取平均后的目标函数，可以看作权重取值构成的高维空间中的一片丘陵地形。负梯度向量指出这片地形中下降最快的方向，使权重趋近某个极小值点，在那里平均输出误差较低。

实际中，大多数从业者使用一种称为随机梯度下降（SGD）的过程。它先输入少量样本的输入向量，计算输出和误差，再计算这些样本的平均梯度，并据此调整权重。对训练集中的许多小批样本反复执行这一过程，直到目标函数的平均值不再下降。之所以称为“随机”，是因为每一小批样本只能给出全部样本平均梯度的一个带噪声估计。与复杂得多的优化技术相比，这个简单过程通常能以令人惊讶的速度找到一组好的权重[18]。训练完成后，在另一组称为测试集的样本上评估系统性能。这用于检验机器的泛化能力，即对训练时从未见过的新输入给出合理答案的能力。

当前机器学习的许多实际应用，仍是在人工设计的特征之上使用线性分类器。二分类线性分类器计算特征向量各分量的加权和。如果该加权和超过一个阈值，就把输入归为某个特定类别。

从 20 世纪 60 年代起，我们就知道线性分类器只能把输入空间划分为非常简单的区域，即由一个超平面分隔的半空间[19]。然而，图像和语音识别等问题要求输入—输出函数对输入中的无关变化不敏感，例如对象位置、方向或光照的变化，以及语音音高或口音的变化；同时又必须对某些细微变化十分敏感，例如白狼与一种外形像狼的白色犬种——萨摩耶犬——之间的差异。在像素层面，两只姿态不同、所处环境不同的萨摩耶犬图像可能相差很大；而姿态相同、背景相似的一只萨摩耶犬和一只狼的图像，却可能非常相似。直接作用于原始像素的线性分类器，或任何其他“浅层”分类器，都不可能既区分后两张图像，又把前两张归入同一类别。因此，浅层分类器需要一个能够解决“选择性—不变性”两难问题的好特征提取器：产生的表示要对图像中有助于判别的方面具有选择性，却对动物姿态等无关方面保持不变。为了增强分类器，可以像核方法[20]那样采用通用的非线性特征，但高斯核等产生的通用特征，不能使学习器很好地泛化到远离训练样本的区域[21]。传统选择是人工设计好的特征提取器，这需要相当多的工程技能和领域知识。但如果能够使用通用学习过程自动学到好特征，这些工作就都可以避免。这正是深度学习的关键优势。

深度学习架构是由简单模块逐层堆叠而成的，其中全部或大多数模块都需要学习，许多模块计算的是非线性输入—输出映射。每个模块变换输入，以同时增强表示的选择性和不变性。采用多个非线性层——例如深度为 5 至 20 层——系统就能实现极为复杂的输入函数：既对细微差别敏感，能够区分萨摩耶犬与白狼，又不受背景、姿态、光照和周围物体等较大无关变化的影响。

用反向传播训练多层架构

从模式识别研究的早期[22,23]开始，研究者就希望用可训练的多层网络替代人工设计的特征。然而，尽管解决方案很简单，直到 20 世纪 80 年代中期，它才被广泛理解。事实证明，简单的随机梯度下降就可以训练多层架构。只要模块相对于其输入和内部权重的函数足够平滑，就能使用反向传播过程计算梯度。20 世纪 70 年代和 80 年代，多个研究团队独立发现了这种做法及其有效性[24 - 27]。

反向传播用来计算目标函数相对于多层模块堆叠中各权重的梯度，本质上只是求导链式法则的一种实用应用。其关键认识在于：从目标函数相对于某模块输出（也就是后续模块输入）的梯度出发，向后计算，就能得到目标函数相对于该模块输入的导数或梯度（图 1）。反向传播方程可以反复应用，把梯度从顶部输出端——网络产生预测的位置——一直传递到底部——接收外部输入的位置。得到这些梯度后，就很容易计算目标函数相对于每个模块权重的梯度。

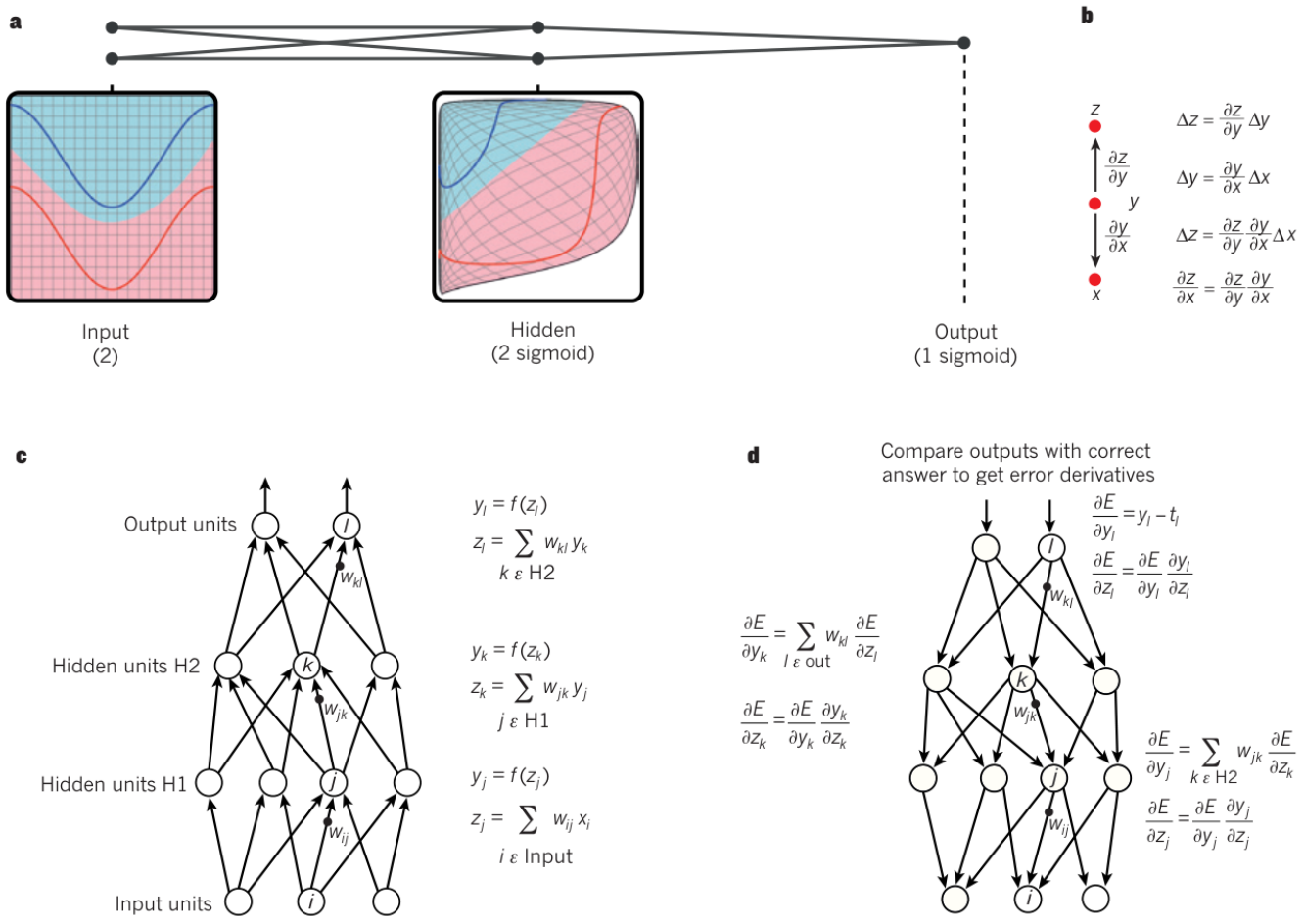

图 1 | 多层神经网络与反向传播。

a，多层神经网络（由相连的点表示）能够扭曲输入空间，使不同类别的数据（红线和蓝线上的样本）变为线性可分。注意，输入空间中的规则网格（左图）也被隐藏单元变换了（中图）。这只是一个包含 2 个输入单元、2 个隐藏单元和 1 个输出单元的示例；用于对象识别或自然语言处理的网络则包含数万或数十万个单元。经 C. Olah 授权转载（http://colah.github.io/）。

b，求导链式法则说明两种微小作用如何复合：x 的微小变化对 y 的影响，以及 y 对 z 的影响。x 的微小变化 Δx 首先乘以 ∂y/∂x，变成 y 的微小变化 Δy，这就是偏导数的定义。类似地，Δy 导致 z 产生变化 Δz。将一个方程代入另一个方程，便得到求导链式法则：Δx 乘以 ∂y/∂x 与 ∂z/∂y 的乘积后变成 Δz。当 x、y、z 是向量、导数是雅可比矩阵时，该法则同样成立。

c，具有 2 个隐藏层和 1 个输出层的神经网络前向计算所用方程。每一层构成一个可以反向传播梯度的模块。在每层中，首先计算每个单元的总输入 z，即下一层各单元输出的加权和；然后将非线性函数 f(•) 作用于 z，得到该单元输出。为简洁起见，图中省略了偏置项。神经网络使用的非线性函数包括近年来常用的整流线性单元（ReLU）f(z)=max(0,z)，以及更传统的 S 形函数，例如双曲正切 f(z)=(exp(z)-exp(-z))/(exp(z)+exp(-z)) 和逻辑函数 f(z)=1/(1+exp(-z))。

d，反向计算所用方程。在每个隐藏层，我们计算误差相对于各单元输出的导数，它是误差相对于上一层各单元总输入的导数的加权和。然后乘以 f(z) 的导数，将误差对输出的导数转为对输入的导数。在输出层，通过对代价函数求导，得到误差相对于单元输出的导数。若单元 1 的代价函数为 $0.5(y_1-t_1)^2$，其中 t_1 是目标值，则该导数为 y_1-t_1。一旦知道 $\partial E/\partial z_k$，从下一层单元 j 连到本层单元 k 的连接权重 w_{jk} 的误差导数就是 $y_j \cdot \partial E/\partial z_k$。

图内术语：Input 输入；Hidden 隐藏层；Output 输出；Input units 输入单元；Hidden units 隐藏单元；Output units 输出单元；Compare outputs with correct answer to get error derivatives 将输出与正确答案比较，得到误差导数。原式保留在图中。译注：图1b英文图注将链式法则乘积的第二项印为 ∂z/∂x；这里依据同图公式写为 ∂z/∂y，英文对照保留原文。

深度学习的许多应用采用前馈神经网络架构（图 1），学习从固定大小的输入（例如一张图像）到固定大小的输出（例如若干类别各自的概率）的映射。从一层到下一层时，一组单元计算来自前一层输入的加权和，并让结果通过一个非线性函数。目前最流行的非线性函数是整流线性单元（ReLU），即半波整流器 f(z)=max(z,0)。过去几十年，神经网络常使用更平滑的非线性函数，如 tanh(z) 或 1/(1+exp(-z))；但在层数很多的网络中，ReLU 通常学习得快得多，从而可以在不进行无监督预训练的情况下训练深度监督网络[28]。不属于输入层或输出层的单元通常称为隐藏单元。隐藏层可以看作对输入进行非线性扭曲，使各类别到最后一层时变得线性可分（图 1）。

20 世纪 90 年代后期，机器学习研究界基本放弃了神经网络和反向传播，计算机视觉与语音识别研究界也忽视了它们。当时普遍认为，在先验知识很少的情况下，无法学到有用的多级特征提取器。尤其是，人们通常认为简单的梯度下降会陷入差的局部极小值，也就是无论怎样微小地改变权重，都不能降低平均误差的权重配置。

实际中，对于大型网络，差的局部极小值很少构成问题。不管初始条件如何，系统几乎总能得到质量非常接近的解。近期理论与实证结果强烈表明，一般来说局部极小值并不是严重问题。相反，目标函数地形中密布着数量呈组合级增长的鞍点：这些点的梯度为零，曲面在大多数维度上向上弯曲，在其余维度上向下弯曲[29,30]。分析似乎表明，只有少数向下弯曲方向的鞍点数量极多，但其中几乎所有点的目标函数值都非常接近。因此，算法具体停在哪一个鞍点上，并不是很重要。

2006 年前后，由加拿大高级研究院（CIFAR）聚集的一批研究者重新激发了人们对深度前馈网络的兴趣[31 – 34]。他们提出了无需带标签数据就能建立多层特征检测器的无监督学习过程。学习每一层特征检测器的目标，是能够重建或建模其下一层特征检测器的活动，或者原始输入。以这一重建目标逐层“预训练”越来越复杂的特征检测器，可以把深度网络的权重初始化到合理值。然后，在网络顶部添加最终输出层，再用标准反向传播微调整个深度系统[33 – 35]。这种方法在手写数字识别和行人检测中效果很好，尤其是在带标签数据极其有限时[36]。

这种预训练方法的第一个重要应用是语音识别。它得益于易于编程的高速图形处理器（GPU）的出现[37]，使研究者能够以原来 10 倍或 20 倍的速度训练网络。2009 年，该方法被用于把从声波中提取的一小段时间窗内的系数，映射为一组概率，表示时间窗中央帧可能对应的各种语音片段。它在一个小词汇量的标准语音识别基准上取得了破纪录的结果[38]，随后很快发展到在大词汇量任务上也刷新纪录[39]。到 2012 年，许多主要语音研究团队都在开发 2009 年这一深度网络的不同版本[6]，并且已将其部署到 Android 手机。对于较小的数据集，无监督预训练有助于防止过拟合[40]；因此，当带标签样本很少，或者在某些

“源”任务样本很多而某些“目标”任务样本很少的迁移场景中，它能显著改善泛化。深度学习重新获得认可后，人们发现，只有小数据集才需要预训练阶段。

不过，有一种特殊的深度前馈网络，比相邻层全连接的网络更容易训练，泛化也好得多。这就是卷积神经网络（ConvNet）[41,42]。在神经网络不受青睐的时期，它已取得许多实际成功，最近又被计算机视觉研究界广泛采用。

卷积神经网络

卷积网络用于处理以多个数组形式表示的数据，例如彩色图像由 3 个二维数组组成，分别存储三个颜色通道中的像素强度。许多模态的数据都呈现为多个数组：信号和序列（包括语言）是一维的，图像或音频频谱图是二维的，视频或体图像则是三维的。卷积网络有四个利用自然信号特性的核心思想：局部连接、权重共享、池化，以及使用多层结构。

典型卷积网络的架构（图 2）由一系列阶段组成。前几个阶段包含两种层：卷积层和池化层。卷积层中的单元组织为特征图；每个单元通过一组称为滤波器组的权重，与前一层特征图中的局部区域相连。这个局部加权和再通过 ReLU 等非线性函数。同一张特征图中的所有单元共享同一滤波器组，同层的不同特征图则使用不同滤波器组。这一架构有两方面原因。首先，在图像等数组数据中，局部的一组数值往往高度相关，会形成容易检测的、有辨识度的局部模式。其次，图像和其他信号的局部统计特性不随位置改变。换言之，一个模式若能出现在图像的某一部分，也可能出现在任何地方。因此，让不同位置的单元共享相同权重，就能在数组不同部分检测相同模式。从数学上说，一张特征图执行的滤波操作是离散卷积，“卷积网络”由此得名。

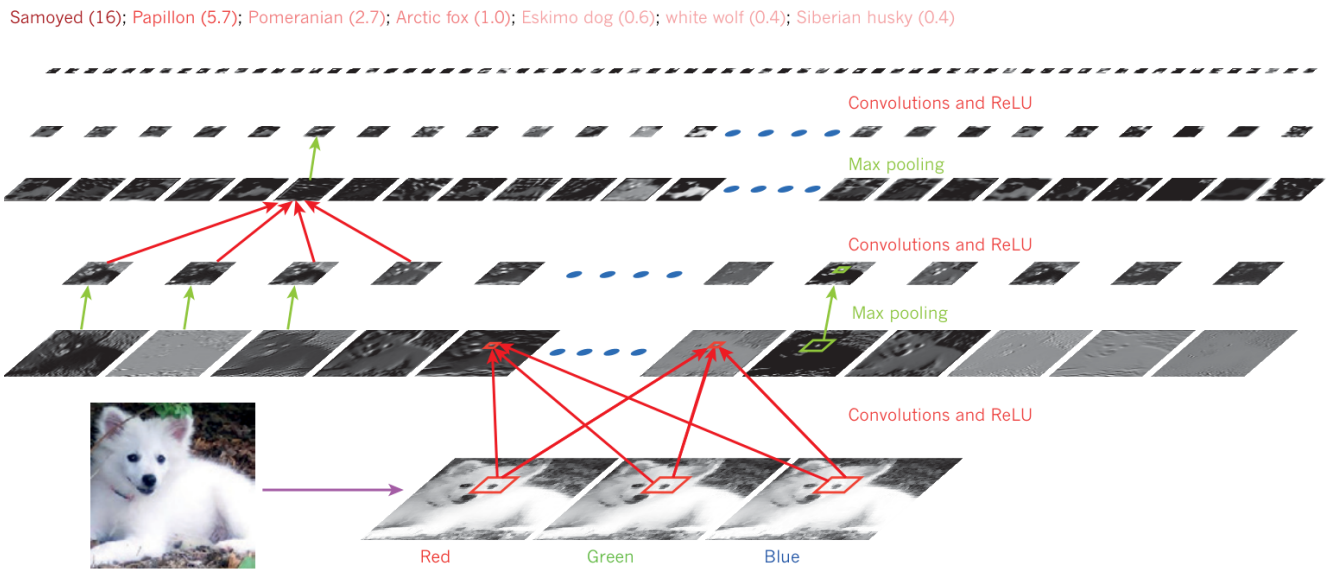

图 2 | 卷积网络内部。图中展示典型卷积网络处理一张萨摩耶犬图像时，各层的输出而非滤波器；每层的输出沿水平方向排列。左下方是输入图像，右下方是 RGB（红、绿、蓝）通道输入。每个矩形图像都是一张特征图，对应某个已学得特征在图像各位置的检测输出。信息自下而上流动；低层特征起有方向性的边缘检测器的作用，输出端为每个图像类别计算分数。ReLU 表示整流线性单元。

图内术语：Convolutions and ReLU 卷积与整流线性单元；Max pooling 最大池化；Red/Green/Blue 红/绿/蓝。输出类别及分数：萨摩耶犬 16；蝴蝶犬 5.7；博美犬 2.7；北极狐 1.0；爱斯基摩犬 0.6；白狼 0.4；西伯利亚哈士奇 0.4。

卷积层的作用是检测前一层特征在局部的组合，而池化层的作用则是将语义上相似的特征合并。由于构成局部模式的各特征，其相对位置可能稍有变化，因此，对各特征的位置进行粗粒化，就能可靠地检测该模

式。典型池化单元计算一张或少数几张特征图中一个局部单元区域的最大值。相邻池化单元所接收的区域相互错开不止一行或一列，因此可以降低表示的维度，并形成对微小平移和形变的不变性。将卷积、非线性变换与池化堆叠两三个阶段后，再接更多卷积层和全连接层。在卷积网络中反向传播梯度，与在普通深度网络中一样简单，因此能够训练所有滤波器组的全部权重。

深度神经网络利用了许多自然信号具有组合式层级结构这一性质：高层特征由低层特征组合而成。图像中，边缘的局部组合形成局部模式，局部模式组成部件，部件再组成对象。语音和文本中也存在类似层级，从声音到音素的具体发音、音位、音节、单词，再到句子。池化使前一层元素的位置和外观发生变化时，表示仍然只发生很小的变化。

卷积网络中的卷积层和池化层，直接受到视觉神经科学中简单细胞、复杂细胞经典概念的启发[43]，其整体架构也类似于视觉皮层腹侧通路中的 LGN－V1－V2－V4－IT 层级[44]。当向卷积网络模型和猴子展示同一幅图片时，卷积网络高层单元的激活能够解释猴颞下皮层中随机选取的 160 个神经元群体反应方差的一半[45]。卷积网络源自新认知机（neocognitron）[46]；后者的架构有些相似，却没有反向传播这样的端到端监督学习算法。一种称为时延神经网络的早期一维卷积网络，被用于识别音位和简单单词[47,48]。

卷积网络的众多应用可以追溯到 20 世纪 90 年代初，最早包括用于语音识别的时延神经网络[47]和文档识读[42]。文档识读系统将卷积网络与实现语言约束的概率模型联合训练。到 20 世纪 90 年代后期，该系统已在识读美国超过 10% 的支票。后来，微软部署了多种基于卷积网络的光学字符识别和手写识别系统[49]。20 世纪 90 年代初，人们还试验了使用卷积网络检测自然图像中的人脸、手等对象[50,51]，以及进行人脸识别[52]。

利用深度卷积网络理解图像

自 21 世纪初以来，卷积网络在图像中对象和区域的检测、分割与识别方面取得了巨大成功。这些任务的带标签数据都比较充足，例如交通标志识别[53]、生物图像分割[54]——特别是连接组学中的分割[55]——以及自然图像中人脸、文字、行人和人体的检测[36,50,51,56－58]。近期卷积网络一个重要的实际成功是人脸识别[59]。

重要的是，图像可以在像素层面进行标注，这将应用于自主移动机器人和自动驾驶汽车等技术[60,61]。Mobileye、NVIDIA 等公司正在其即将推出的车载视觉系统中使用这种基于卷积网络的方法。另一些日益重要的应用包括自然语言理解[14]和语音识别[7]。

尽管有这些成功，直到 2012 年 ImageNet 竞赛之前，主流计算机视觉和机器学习研究界仍基本忽视卷积网络。将深度卷积网络用于一个约含 100 万张网络图像、涵盖 1,000 个类别的数据集时，它取得了惊人的结果，使最强竞争方法的错误率几乎减半[1]。这一成功来自对 GPU 的高效使用、ReLU、一种称为 dropout 的新正则化技术[62]，以及通过使已有样本发生形变来产生更多训练样本的技术。它在计算机视觉领域引发了一场革命；卷积网络如今成为几乎所有识别和检测任务的主导方法[4,58,59,63－65]，并在某些任务上接近人类表现。近期一项令人瞩目的展示，将卷积网络和循环网络模块结合起来，生成图像描述（图 3）。

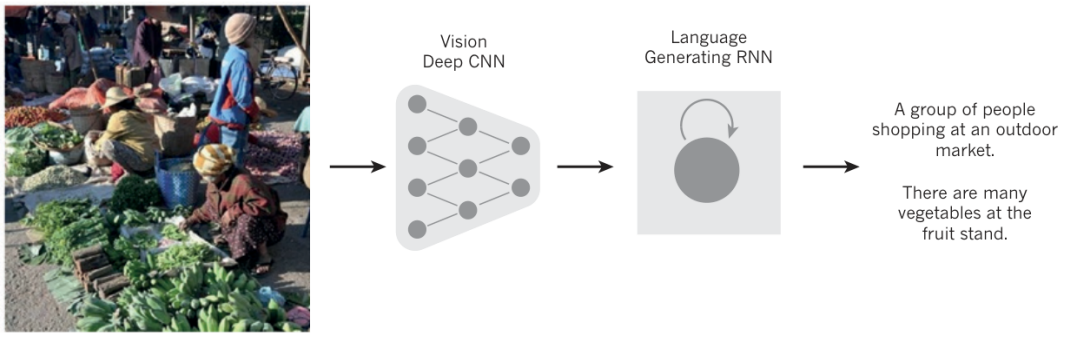



图 3 | 从图像到文本。循环神经网络（RNN）将深度卷积神经网络（CNN）从测试图像提取的表示作为额外输入，并生成描述；训练 RNN 的目标是把图像的高层表示“翻译”为描述文字（上图）。经文献[102]授权转载。当赋予 RNN 在生成每个词（粗体）时将注意力聚焦到输入图像不同位置的能力时（中、下图；较亮区域得到更多注意力），我们发现[86]，它会利用这一能力，更好地把图像“翻译”为描述。

图中生成文字的中文对应：一群人在露天市场购物；水果摊上有许多蔬菜；一位女士正在公园里扔飞盘；一只狗站在硬木地板上；一个停车标志立在道路上，背景是山；一个小女孩坐在床上，身旁有一只泰迪熊；一群人坐在水中的一艘船上；一只长颈鹿站在森林里，背景是树木。Vision/Deep CNN 表示视觉/深度卷积神经网络；Language/Generating RNN 表示语言/生成式循环神经网络。

近期的卷积网络架构有 10 至 20 层 ReLU、数亿个权重，以及单元之间数十亿条连接。仅在两年前，训练如此庞大的网络可能还需要数周；硬件、软件和算法并行化的进步，已将训练时间缩短到数小时。

基于卷积网络的视觉系统所展现的性能，促使 Google、Facebook、Microsoft、IBM、Yahoo!、Twitter、Adobe 等多数主要科技公司，以及迅速增多的初创企业，启动研究开发项目，并部署基于卷积网络的图像理解产品和服务。

卷积网络很容易在芯片或现场可编程门阵列上实现高效硬件运算[66,67]。NVIDIA、Mobileye、Intel、Qualcomm、Samsung 等公司正在开发卷积网络芯片，以支持智能手机、相机、机器人和自动驾驶汽车上的实时视觉应用。

分布式表示与语言处理

深度学习理论表明，与不使用分布式表示的经典学习算法相比，深度网络具有两种不同的指数级优势[21]。这两种优势均源于组合的力量，并依赖于底层数据生成分布具有合适的组合结构[40]。第一，学习分布式表示能够泛化到训练时未出现过的、已学特征取值的新组合，例如 n 个二值特征可以构成 2^n 种组合[68,69]。第二，在深度网络中组合多层表示，还有可能带来另一种随深度呈指数增长的优势[70]。

多层神经网络的隐藏层会学习一种输入表示，使目标输出易于预测。一个很好的例子是：训练多层神经网络，根据前面若干词组成的局部上下文，预测序列中的下一个词[71]。上下文中的每个词以 N 选一向量输入网络，即一个分量为 1，其余为 0。在第一层，每个词产生不同的激活模式，也就是词向量（图 4）。语言模型中，网络的其他层学习把输入词向量变换为预测下一个词的输出词向量，由此可预测词表中任意词成为下一个词的概率。网络学到的词向量含有许多活跃分量，每个分量都可以解释为该词的一个单独的特征；这一点最早在学习符号的分布式表示时得到展示[27]。这些语义特征并没有显式存在于输入中，而是由学习过程发现的：它们把输入、输出符号之间的结构性关系分解为多条“微规则”。后来发现，当词序列来自大型真实文本语料、单条微规则不可靠时，词向量学习仍然非常有效[71]。例如，在预测新闻报道中下一个词的训练中，Tuesday（星期二）与 Wednesday（星期三）的词向量会非常相似，Sweden（瑞典）与 Norway（挪威）也一样。这类表示称为分布式表示，因为其元素（特征）并不相互排斥，众多配置对应于观测数据中的不同变化。组成这些词向量的特征，不是由专家事先指定，而是神经网络自动发现的。如今，从文本中学习到的词向量表示已广泛用于自然语言应用[14,17,72–76]。

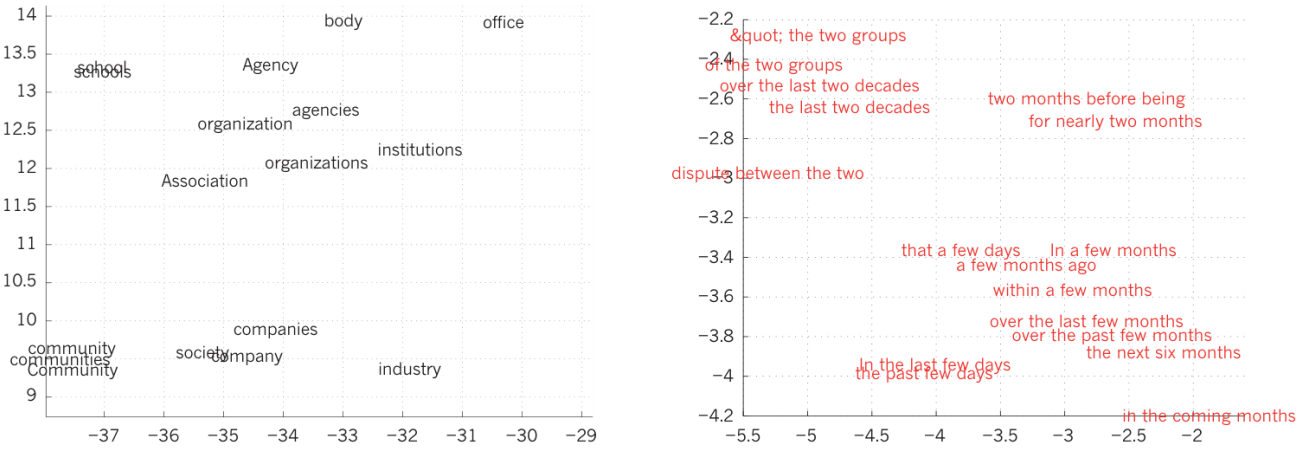

图 4 | 已学习词向量的可视化。左图展示为语言建模学得的词表示，使用 t-SNE 算法[103]将其非线性投影到二维以便可视化。右图是由英译法编码器—解码器循环神经网络学习到的短语的二维表示[75]。可以观察到，语义相似的词或词序列被映射到相邻的表示。词的分布式表示通过反向传播获得：同时学习每个词的表示，以及预测某个目标量的函数，例如语言建模中的下一个词，或机器翻译中的整个译文词序列[18,75]。

图内坐标为二维可视化坐标。左图示例包括 community（社区）、society（社会）、industry（行业）、company（公司）、school（学校）、organization（组织）、institution（机构）、office（办公室）、agency（机构）、association（协会）等及其复数形式。右图展示有关过去/未来数月、数日、两个群体和过去二十年等短语的表示，原短语保留以便比较其语义邻近关系。

表示问题处于认知研究中逻辑启发范式与神经网络启发范式之争的核心。在逻辑启发范式中，一个符号实例唯一的属性，是它与其他符号实例相同或不同；它没有与使用方式有关的内部结构。要用符号推理，必须把它们绑定到精心选定的推理规则中的变量上。相反，神经网络仅用大型激活向量、大型权重矩阵和标量非线性函数，就能执行那种支撑轻松常识推理的快速“直觉式”推断。

在神经语言模型提出之前[71]，语言统计建模的标准方法并不利用分布式表示，而是统计长度不超过 N 的短符号序列（称为 N 元语法）的出现频率。可能的 N 元语法数量的量级为 V^N，其中 V 是词表大小，因此，要考虑超过寥寥几个词的上下文，就需要极大的训练语料。N 元语法把每个词当作一个原子单元，因此无法在语义相关的词序列之间进行泛化；神经语言模型则能够做到，因为它为每个词赋予一个由实值特征组成的向量，语义相关的词最终会在该向量空间中相互靠近（图 4）。

循环神经网络

反向传播刚提出时，最令人兴奋的用途是训练循环神经网络（RNN）。对于语音、语言等涉及序列输入的任务，使用 RNN 往往更好（图 5）。RNN 每次处理输入序列中的一个元素，并在隐藏单元中维持一个“状态向量”，隐式包含序列中全部过去元素的历史信息。如果把不同离散时间步的隐藏单元输出，看作深层多层网络中不同神经元的输出（图 5 右图），就很容易理解如何运用反向传播训练 RNN。

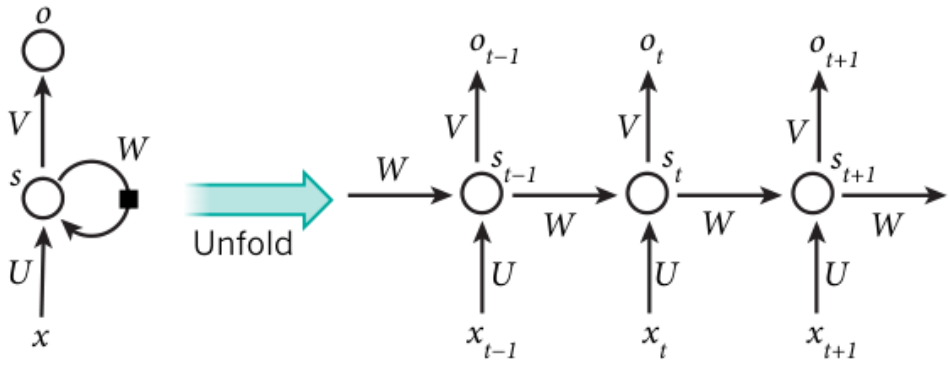

图 5 | 循环神经网络及其前向计算沿时间的展开。人工神经元，例如归于节点 s、在时刻 t 取值为 s_t 的隐藏单元，接收来自先前时间步其他神经元的输入（左图中的黑色方块代表一个时间步的延迟）。这样，循环神经网络就可以把由元素 x_t 构成的输入序列映射为由元素 o_t 构成的输出序列，每个 o_t 都依赖此前所有 $x_{t'}$，其中 $t' \le t$。每个时间步使用相同的参数矩阵 U、V、W。还存在许多其他架构，例如一种变体可以生成输出序列（如单词），并把每次输出用作下一时间步的输入。反向传播算法（图 1）可以直接应用于右图展开网络的计算图，以计算总误差——例如生成正确输出序列的对数概率——相对于所有状态 s_t 和全部参数的导数。

图内术语：Unfold 沿时间展开；x 为输入，s 为状态，o 为输出；U、V、W 为各时间步共享的参数矩阵。

RNN 是非常强大的动力系统，但事实证明，其训练存在困难：反向传播的梯度在每个时间步都会变大或变小，因此跨越很多时间步后，通常会发生梯度爆炸或梯度消失[77,78]。

得益于网络架构[79,80]和训练方式[81,82]的进步，人们发现 RNN 非常擅长预测文本中的下一个字符[83]或序列中的下一个词[75]；它还可以处理更复杂的任务。例如，逐词读入一句英语后，可以训练一个英语“编码器”网络，使隐藏单元的最终状态向量很好地表示该句表达的思想。这个思想向量可作为联合训练的法语“解码器”网络的初始隐藏状态，或额外输入；后者输出法语译文第一个词的概率分布。从该分布选择某个首词并输入解码器后，它就会输出第二个词的概率分布，如此继续，直到选出句号[17,72,76]。总体而言，该过程按照一个依赖英语句子的概率分布生成法语词序列。这种相当朴素的机器翻译方法，很快便能与当时最先进的方法竞争，由此让人强烈怀疑：理解句子是否真的需要某种用推理规则操纵的内部符号表达式。它更符合另一种观点，即日常推理涉及许多同时进行的类比，各自为结论增添可信性[84,85]。

除了把法语句子的含义翻译成英语句子，还可以学习把图像的含义“翻译”为英语句子（图 3）。这里的编码器是深度卷积网络，把像素变换成最后一个隐藏层中的激活向量。解码器则是与机器翻译、神经语言建模所用网络相似的 RNN。近期，这类系统引发了大量关注，实例见文献[86]。

RNN 沿时间展开后（图 5），可以看作各层共享同一组权重的极深前馈网络。尽管其主要目的是学习长期依赖，理论与实证证据表明，要学会长时间保存信息仍然很困难[78]。

为解决这一问题，一种思路是给网络增加显式记忆。此类方法中最早提出的是长短期记忆（LSTM）网络，使用本身就适于长时间记住输入的特殊隐藏单元[79]。一种称为记忆单元的特殊单元，类似于累加器或带门控的泄漏神经元：它有一条通向下一时间步自身的连接，权重为 1，因此能够复制自身实值状态并累积外部信号；但该自连接受到另一个单元的乘性门控，后者学习决定何时清除记忆内容。

后来证明，LSTM 网络比传统 RNN 更有效，尤其是每个时间步包含多层时[87]。它使完整的语音识别系统能够直接从声学输入得到转写文本的字符序列。目前，LSTM 或相关门控单元也被用于机器翻译中表现优异的编码器、解码器网络[17,72,76]。

过去一年，多位作者提出了给 RNN 增加记忆模块的不同方案。其中包括神经图灵机：它给网络增加“纸带式”记忆，RNN 可以选择读取或写入[88]；以及记忆网络：给普通网络增加一种联想记忆[89]。记忆网络在标准问答基准上取得了出色表现。记忆用于保存故事，之后网络再回答有关该故事的问题。

除简单记忆外，神经图灵机和记忆网络还被用于通常需要推理和符号操作的任务。神经图灵机可以被教授“算法”。例如，输入是一个未排序序列，每个符号附带一个表示其在列表中优先级的实数时，它能够学习输出排好序的符号列表[88]。记忆网络可以在类似文字冒险游戏的环境中，经过训练来追踪世界状态；读完一个故事后，它可以回答需要复杂推断的问题[90]。在一个测试例子中，网络读入由 15 个句子组成的《魔戒》故事，并正确回答“佛罗多现在在哪里？”等问题[89]。

深度学习的未来

无监督学习[91–98]曾在重新激发深度学习研究兴趣方面起到催化作用，但后来其光芒被纯监督学习的成功所掩盖。虽然本综述没有重点讨论无监督学习，我们预计，从更长远看，它会变得重要得多。人类和动物的学习大多是无监督的：我们通过观察发现世界的结构，而不是依靠别人告诉我们每一个对象的名称。

人类视觉是一个主动过程：利用面积小、分辨率高的中央凹和范围大、分辨率低的周边视野，以智能且针对具体任务的方式，依次采样视野。我们预计，未来视觉领域的许多进展，将来自经过端到端训练、将卷积网络与 RNN 结合，并利用强化学习决定看向何处的系统。深度学习与强化学习相结合的系统尚处于起步阶段，但它们在分类任务上已经超过被动视觉系统[99]，在学习玩多种电子游戏方面也取得了令人印象深刻的结果[100]。

自然语言理解是深度学习有望在未来几年产生重大影响的另一个领域。我们预计，使用 RNN 理解句子或整篇文档的系统，在学会每次选择性地关注某一部分的策略后，表现将大幅改善[76,86]。

最终，人工智能的重大进展将来自把表示学习与复杂推理结合起来的系统。尽管深度学习和简单推理早已用于语音及手写识别，仍然需要新的范式，用对大型向量的操作，替代基于规则的符号表达式操纵[101]。

收稿：2015 年 2 月 25 日；接受：2015 年 5 月 1 日。

致谢

作者感谢加拿大自然科学与工程研究理事会、加拿大高级研究院（CIFAR）、美国国家科学基金会和美国海军研究办公室提供支持。Y.L. 与 Y.B. 为 CIFAR 研究员。

作者信息

转载与授权信息见 www.nature.com/reprints。作者声明不存在竞争性财务利益。欢迎读者在本文网络版 go.nature.com/7cjbaa 留言。通信请联系 Y.L.（yann@cs.nyu.edu）。

参考文献

书目信息保留原文；原作者对重点文献的评述置于相应条目之后，以中文呈现。

[1] Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012).

这项突破性研究使用卷积网络，将对象识别错误率几乎减半，并促使计算机视觉研究界迅速采用深度学习。

[2] Farabet, C., Couprie, C., Najman, L. & LeCun, Y. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929 (2013).

[3] Tompson, J., Jain, A., LeCun, Y. & Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In Proc. Advances in Neural Information Processing Systems 27 1799–1807 (2014).

[4] Szegedy, C. et al. Going deeper with convolutions. Preprint at http://arxiv.org/ abs/1409.4842 (2014).

[5] Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocky, J. Strategies for training large scale neural network language models. In Proc. Automatic Speech Recognition and Understanding 196–201 (2011).

[6] Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine 29, 82–97 (2012).

这篇由主要语音识别实验室联合发表的论文，总结了深度学习在自动语音识别的音位分类任务上取得的突破，这是深度学习的首个重要工业应用。

[7] Sainath, T., Mohamed, A.-R., Kingsbury, B. & Ramabhadran, B. Deep convolutional neural networks for LVCSR. In Proc. Acoustics, Speech and Signal Processing 8614–8618 (2013).

[8] Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. 55, 263–274 (2015).

[9] Ciodaro, T., Deva, D., de Seixas, J. & Damazio, D. Online particle detection with neural networks based on topological calorimetry information. J. Phys. Conf. Series 368, 012030 (2012).

[10] Kaggle. Higgs boson machine learning challenge. Kaggle https://www.kaggle. com/c/higgs-boson (2014).

[11] Helmstaedter, M. et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500, 168–174 (2013).

[12] Leung, M. K., Xiong, H. Y., Lee, L. J. & Frey, B. J. Deep learning of the tissue-regulated splicing code. Bioinformatics 30, i121–i129 (2014).

[13] Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 6218 (2015).

[14] Collobert, R., et al. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537 (2011).

[15] Bordes, A., Chopra, S. & Weston, J. Question answering with subgraph embeddings. In Proc. Empirical Methods in Natural Language Processing http:// arxiv.org/abs/1406.3676v3 (2014).

[16] Jean, S., Cho, K., Memisevic, R. & Bengio, Y. On using very large target vocabulary for neural machine translation. In Proc. ACL-IJCNLP http://arxiv.org/ abs/1412.2007 (2015).

[17] Sutskever, I. Vinyals, O. & Le. Q. V. Sequence to sequence learning with neural networks. In Proc. Advances in Neural Information Processing Systems 27 3104–3112 (2014).

本文利用文献[72]提出的架构，取得了当时最先进的机器翻译结果：训练循环网络读取一种语言的句子，产生其含义的语义表示，再生成另一种语言的译文。

[18] Bottou, L. & Bousquet, O. The tradeoffs of large scale learning. In Proc. Advances in Neural Information Processing Systems 20 161–168 (2007).

[19] Duda, R. O. & Hart, P. E. Pattern Classification and Scene Analysis (Wiley, 1973).

[20] Schölkopf, B. & Smola, A. Learning with Kernels (MIT Press, 2002).

[21] Bengio, Y., Delalleau, O. & Le Roux, N. The curse of highly variable functions for local kernel machines. In Proc. Advances in Neural Information Processing Systems 18 107–114 (2005).

[22] Selfridge, O. G. Pandemonium: a paradigm for learning in mechanisation of thought processes. In Proc. Symposium on Mechanisation of Thought Processes 513–526 (1958).

[23] Rosenblatt, F. The Perceptron — A Perceiving and Recognizing Automaton. Tech. Rep. 85-460-1 (Cornell Aeronautical Laboratory, 1957).

[24] Werbos, P. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard Univ. (1974).

[25] Parker, D. B. Learning Logic Report TR–47 (MIT Press, 1985).

[26] LeCun, Y. Une procédure d'apprentissage pour Réseau à seuil assymétrique in Cognitiva 85: a la Frontière de l'Intelligence Artificielle, des Sciences de la Connaissance et des Neurosciences [in French] 599–604 (1985).

[27] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. Nature 323, 533–536 (1986).

[28] Glorot, X., Bordes, A. & Bengio. Y. Deep sparse rectifier neural networks. In Proc. 14th International Conference on Artificial Intelligence and Statistics 315–323 (2011).

本文表明，如果隐藏层由 ReLU 构成，对极深神经网络进行监督训练的速度会快得多。

[29] Dauphin, Y. et al. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Proc. Advances in Neural Information Processing Systems 27 2933–2941 (2014).

[30] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B. & LeCun, Y. The loss surface of multilayer networks. In Proc. Conference on AI and Statistics http:// arxiv.org/abs/1412.0233 (2014).

[31] Hinton, G. E. What kind of graphical model is the brain? In Proc. 19th International Joint Conference on Artificial intelligence 1765–1775 (2005).

[32] Hinton, G. E., Osindero, S. & Teh, Y.-W. A fast learning algorithm for deep belief nets. Neural Comp. 18, 1527–1554 (2006).

本文提出一种新颖而有效的极深神经网络训练方法：采用受限玻尔兹曼机的无监督学习过程，每次预训练一个隐藏层。

[33] Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H. Greedy layer-wise training of deep networks. In Proc. Advances in Neural Information Processing Systems 19 153–160 (2006).

本报告表明，文献[32]提出的无监督预训练方法显著提高了测试数据上的性能，并将其推广到自编码器等其他无监督表示学习技术。

[34] Ranzato, M., Poultney, C., Chopra, S. & LeCun, Y. Efficient learning of sparse representations with an energy-based model. In Proc. Advances in Neural Information Processing Systems 19 1137–1144 (2006).

[35] Hinton, G. E. & Salakhutdinov, R. Reducing the dimensionality of data with neural networks. Science 313, 504–507 (2006).

[36] Sermanet, P., Kavukcuoglu, K., Chintala, S. & LeCun, Y. Pedestrian detection with unsupervised multi-stage feature learning. In Proc. International Conference on Computer Vision and Pattern Recognition http://arxiv.org/abs/1212.0142 (2013).

[37] Raina, R., Madhavan, A. & Ng, A. Y. Large-scale deep unsupervised learning using graphics processors. In Proc. 26th Annual International Conference on Machine Learning 873–880 (2009).

[38] Mohamed, A.-R., Dahl, G. E. & Hinton, G. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20, 14–22 (2012).

[39] Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

[40] Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Machine Intell. 35, 1798–1828 (2013).

[41] LeCun, Y. et al. Handwritten digit recognition with a back-propagation network. In Proc. Advances in Neural Information Processing Systems 396–404 (1990).

这是第一篇使用反向传播训练卷积网络、以分类低分辨率手写数字图像的论文。

[42] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324 (1998).

这篇综述讨论使用基于梯度的优化，端到端训练深度神经网络等模块化系统的原理，并展示如何将神经网络，尤其是卷积网络，与搜索或推断机制结合起来，对彼此依赖的复杂输出建模，例如与文档内容对应的字符序列。

[43] Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154 (1962).

[44] Felleman, D. J. & Essen, D. C. V. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47 (1991).

[45] Cadieu, C. F. et al. Deep neural networks rival the representation of primate it cortex for core visual object recognition. PLoS Comp. Biol. 10, e1003963 (2014).

[46] Fukushima, K. & Miyake, S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition 15, 455–469 (1982).

[47] Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K. & Lang, K. Phoneme recognition using time-delay neural networks. IEEE Trans. Acoustics Speech Signal Process. 37, 328–339 (1989).

[48] Bottou, L., Fogelman-Soulié, F., Blanchet, P. & Lienard, J. Experiments with time delay networks and dynamic time warping for speaker independent isolated digit recognition. In Proc. EuroSpeech 89 537–540 (1989).

[49] Simard, D., Steinkraus, P. Y. & Platt, J. C. Best practices for convolutional neural networks. In Proc. Document Analysis and Recognition 958–963 (2003).

[50] Vaillant, R., Monrocq, C. & LeCun, Y. Original approach for the localisation of objects in images. In Proc. Vision, Image, and Signal Processing 141, 245–250 (1994).

[51] Nowlan, S. & Platt, J. in Neural Information Processing Systems 901–908 (1995).

[52] Lawrence, S., Giles, C. L., Tsoi, A. C. & Back, A. D. Face recognition: a convolutional neural-network approach. IEEE Trans. Neural Networks 8, 98–113 (1997).

[53] Ciresan, D., Meier, U. Masci, J. & Schmidhuber, J. Multi-column deep neural network for traffic sign classification. Neural Networks 32, 333–338 (2012).

[54] Ning, F. et al. Toward automatic phenotyping of developing embryos from videos. IEEE Trans. Image Process. 14, 1360–1371 (2005).

[55] Turaga, S. C. et al. Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Comput. 22, 511–538 (2010).

[56] Garcia, C. & Delakis, M. Convolutional face finder: a neural architecture for fast and robust face detection. IEEE Trans. Pattern Anal. Machine Intell. 26, 1408–1423 (2004).

[57] Osadchy, M., LeCun, Y. & Miller, M. Synergistic face detection and pose estimation with energy-based models. J. Mach. Learn. Res. 8, 1197–1215 (2007).

[58] Tompson, J., Goroshin, R. R., Jain, A., LeCun, Y. Y. & Bregler, C. C. Efficient object localization using convolutional networks. In Proc. Conference on Computer Vision and Pattern Recognition http://arxiv.org/abs/1411.4280 (2014).

[59] Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Deepface: closing the gap to human-level performance in face verification. In Proc. Conference on Computer Vision and Pattern Recognition 1701–1708 (2014).

[60] Hadsell, R. et al. Learning long-range vision for autonomous off-road driving. J. Field Robot. 26, 120–144 (2009).

[61] Farabet, C., Couprie, C., Najman, L. & LeCun, Y. Scene parsing with multiscale feature learning, purity trees, and optimal covers. In Proc. International Conference on Machine Learning http://arxiv.org/abs/1202.2160 (2012).

[62] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. J. Machine Learning Res. 15, 1929–1958 (2014).

[63] Sermanet, P. et al. Overfeat: integrated recognition, localization and detection using convolutional networks. In Proc. International Conference on Learning Representations http://arxiv.org/abs/1312.6229 (2014).

[64] Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. Conference on Computer Vision and Pattern Recognition 580–587 (2014).

[65] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proc. International Conference on Learning Representations http://arxiv.org/abs/1409.1556 (2014).

[66] Boser, B., Sackinger, E., Bromley, J., LeCun, Y. & Jackel, L. An analog neural network processor with programmable topology. J. Solid State Circuits 26, 2017–2025 (1991).

[67] Farabet, C. et al. Large-scale FPGA-based convolutional networks. In Scaling up Machine Learning: Parallel and Distributed Approaches (eds Bekkerman, R., Bilenko, M. & Langford, J.) 399–419 (Cambridge Univ. Press, 2011).

[68] Bengio, Y. Learning Deep Architectures for AI (Now, 2009).

[69] Montufar, G. & Morton, J. When does a mixture of products contain a product of mixtures? J. Discrete Math. 29, 321–347 (2014).

[70] Montufar, G. F., Pascanu, R., Cho, K. & Bengio, Y. On the number of linear regions of deep neural networks. In Proc. Advances in Neural Information Processing Systems 27 2924–2932 (2014).

[71] Bengio, Y., Ducharme, R. & Vincent, P. A neural probabilistic language model. In Proc. Advances in Neural Information Processing Systems 13 932–938 (2001).

本文提出了神经语言模型，它学习把单词符号转换为由已学语义特征构成的词向量或词嵌入，以预测序列中的下一个词。

[72] Cho, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proc. Conference on Empirical Methods in Natural Language Processing 1724–1734 (2014).

[73] Schwenk, H. Continuous space language models. Computer Speech Lang. 21, 492–518 (2007).

[74] Socher, R., Lin, C. C-Y., Manning, C. & Ng, A. Y. Parsing natural scenes and natural language with recursive neural networks. In Proc. International Conference on Machine Learning 129–136 (2011).

[75] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In Proc. Advances in Neural Information Processing Systems 26 3111–3119 (2013).

[76] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proc. International Conference on Learning Representations http://arxiv.org/abs/1409.0473 (2015).

[77] Hochreiter, S. Untersuchungen zu dynamischen neuronalen Netzen [in German] Diploma thesis, T.U. Münich (1991).

[78] Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Networks 5, 157–166 (1994).

[79] Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. 9, 1735–1780 (1997).

本文提出 LSTM 循环网络。由于它善于学习长距离依赖，已成为近期循环网络进展中的关键组成部分。

[80] ElHihi, S. & Bengio, Y. Hierarchical recurrent neural networks for long-term dependencies. In Proc. Advances in Neural Information Processing Systems 8 http://papers.nips.cc/paper/1102-hierarchical-recurrent-neural-networks-for-long-term-dependencies (1995).

[81] Sutskever, I. Training Recurrent Neural Networks. PhD thesis, Univ. Toronto (2012).

[82] Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. In Proc. 30th International Conference on Machine Learning 1310– 1318 (2013).

[83] Sutskever, I., Martens, J. & Hinton, G. E. Generating text with recurrent neural networks. In Proc. 28th International Conference on Machine Learning 1017– 1024 (2011).

[84] Lakoff, G. & Johnson, M. Metaphors We Live By (Univ. Chicago Press, 2008).

[85] Rogers, T. T. & McClelland, J. L. Semantic Cognition: A Parallel Distributed Processing Approach (MIT Press, 2004).

[86] Xu, K. et al. Show, attend and tell: Neural image caption generation with visual attention. In Proc. International Conference on Learning Representations http:// arxiv.org/abs/1502.03044 (2015).

[87] Graves, A., Mohamed, A.-R. & Hinton, G. Speech recognition with deep recurrent neural networks. In Proc. International Conference on Acoustics, Speech and Signal Processing 6645–6649 (2013).

[88] Graves, A., Wayne, G. & Danihelka, I. Neural Turing machines. http://arxiv.org/ abs/1410.5401 (2014).

[89] Weston, J. Chopra, S. & Bordes, A. Memory networks. http://arxiv.org/ abs/1410.3916 (2014).

[90] Weston, J., Bordes, A., Chopra, S. & Mikolov, T. Towards AI-complete question answering: a set of prerequisite toy tasks. http://arxiv.org/abs/1502.05698 (2015).

[91] Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The wake-sleep algorithm for unsupervised neural networks. Science 268, 1558–1161 (1995).

[92] Salakhutdinov, R. & Hinton, G. Deep Boltzmann machines. In Proc. International Conference on Artificial Intelligence and Statistics 448–455 (2009).

[93] Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In Proc. 25th International Conference on Machine Learning 1096–1103 (2008).

[94] Kavukcuoglu, K. et al. Learning convolutional feature hierarchies for visual recognition. In Proc. Advances in Neural Information Processing Systems 23 1090–1098 (2010).

[95] Gregor, K. & LeCun, Y. Learning fast approximations of sparse coding. In Proc. International Conference on Machine Learning 399–406 (2010).

[96] Ranzato, M., Mnih, V., Susskind, J. M. & Hinton, G. E. Modeling natural images using gated MRFs. IEEE Trans. Pattern Anal. Machine Intell. 35, 2206–2222 (2013).

[97] Bengio, Y., Thibodeau-Laufer, E., Alain, G. & Yosinski, J. Deep generative stochastic networks trainable by backprop. In Proc. 31st International Conference on Machine Learning 226–234 (2014).

[98] Kingma, D., Rezende, D., Mohamed, S. & Welling, M. Semi-supervised learning with deep generative models. In Proc. Advances in Neural Information Processing Systems 27 3581–3589 (2014).

[99] Ba, J., Mnih, V. & Kavukcuoglu, K. Multiple object recognition with visual attention. In Proc. International Conference on Learning Representations http:// arxiv.org/abs/1412.7755 (2014).

[100] Mnih, V. et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015).

[101] Bottou, L. From machine learning to machine reasoning. Mach. Learn. 94, 133–149 (2014).

[102] Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: a neural image caption generator. In Proc. International Conference on Machine Learning http:// arxiv.org/abs/1502.03044 (2014).

[103] van der Maaten, L. & Hinton, G. E. Visualizing data using t-SNE. J. Mach. Learn. Research 9, 2579–2605 (2008).

术语表

English	中文
deep learning	深度学习
representation learning	表示学习
distributed representation	分布式表示
feature map	特征图
backpropagation	反向传播
stochastic gradient descent (SGD)	随机梯度下降
convolutional neural network (ConvNet/CNN)	卷积神经网络
recurrent neural network (RNN)	循环神经网络
rectified linear unit (ReLU)	整流线性单元
pooling	池化
invariance	不变性
saddle point	鞍点
word embedding	词嵌入
long short-term memory (LSTM)	长短期记忆
memory cell	记忆单元
attention	注意力

译制日期：2026-11-15。逐段中英对照、原文页码与来源块编号见同目录“中英对照”。